Research Paper | Statistics | E-ISSN No : 2454-9916 | Volume : 4 | Issue : 11 | Nov 2018

USING HISTORICAL DATA TO INTRODUCE THE STATISTICAL REGRESSION CONCEPT

José Antonio Camúñez-Ruiz [1] | Mª Dolores Pérez-Hidalgo [1]

[1] Facultad de Ciencias Económicas y Empresariales. Universidad de Sevilla. Spain, Avda. Ramón y Cajal, 1, 41018-Sevilla. Spain.

ABSTRACT

To initiate the student in the concept of Linear Regression, and in that of Associated Correlation, in an introductory course of Descriptive Statistics the use of historical problems related to the subject is proposed, in particular, of the problems addressed by Galton at the end of the century XIX, associated with genetic inheritance. In addition to transferring the student to the historical context in which they emerged, giving them names of pioneer scientists, we proceed with the resolution of these problems intuitively, visualizing them graphically, testing with different measures of descriptive statistics already known to them, with trial and error, with the support of calculator and spreadsheet. For the end we leave the formalization of the theory. The experience proposes, then, a reverse course to what is usually habitual, from practice to theory. First the student faces a set of problems in which the base is the relationship between variables, unknown by them until the moment of facing them, and using the statistical instruments related to the management of a variable, such as the mean and the standard deviation. Our role as a teacher is to introduce clues that help you overcome obstacles on a practical level. Then those tracks, formalized, will become the theoretical basis of the chapter devoted to regression analysis. We proceed to the evaluation of the experience. The results of the same, valued in three ways, resolution of practical exercises, survey of students and exam grades, show that the process has been positive for the ultimate goal of learning.

KEYWORDS: Regression, correlation, Galton, Pearson, spreadsheet.

1. INTRODUCTION:

The regression analysis is one of the fundamental pillars in the programs of those subjects that initiate the student in the learning of statistical techniques. It is well known, on the other hand, that these techniques become valuable instruments for the professional development of graduates in Business Administration and Management. The association of the cause-effect type, the relationships between variables, the intensity levels of these relationships, the predictions of behaviors based on a few clues, all constitute a set of ideas that, at a statistical level, is part of the broad concept of "regression". In the initiation courses, the concept is introduced in a descriptive way, handling observed data. Then, once the student starts in the calculation of probabilities and, therefore, in the theoretical probabilistic models, again the regression is incorporated into the programs of the subjects, but now between random variables. In this case, of course, the concepts already acquired at a descriptive level are fundamental. Finally, in the discipline known as Econometrics, which is part of the curriculum of the aforementioned degree in higher education, regression is the basis on which most of the econometric analysis techniques are based. We are aware, therefore, of the importance of this chapter and we are concerned that the absorption of knowledge and ideas is complete and, in addition, that it is attractive. Interesting in this aspect the work of Stanton (2001), which presents a teaching approach from historical data, and that has served as a guide for the development of this experience.

Traditionally, this chapter has been taught following this direction: first, the linear correlation coefficient is introduced, that is, the concept of linear association between variables (with its associated formula) and, then, the regression line with the theoretical calculation of coefficients and measures of goodness of fit. All this accompanied by important and, sometimes, tedious algebraic developments (although necessary), for the justification of the different formulas. The connection between correlation and regression is obscured by occurring more at the algebraic level than at the level of ideas. The theme is completed with a series of practical exercises and problems, first developed by the teacher, and then proposed, where the student has to show the knowledge and skills acquired during the previous explanations, both theoretical and practical, of the teacher. The entire learning process takes a vertical direction, from top to bottom, from the teacher to the student, with little reflection and maturation on the new and important ideas. The vast majority of published manuals on the subject follow that path. Our experience as professors (up to 20 years of teaching) shows us a somewhat somber and unsatisfactory scenario, because this topic is a bit dislocated in that immense ocean of statistics.

2. MATERIALS AND METHODS:

The described thing has made us pose, for some time, the search of educational alternatives. One of them would be (and is the one we have developed here) the same one that served Galton to introduce himself in these concepts at the end of the 19th century. Galton, a cousin of Darwin, and a recognized scientist of that century in his own right, has often been criticized for his commitment to eugenics. On the other hand, there are those who believe that the lasting fame of his cousin has unjustly overshadowed the important scientific contributions with which Galton contributed to the field of biology, psychology and applied statistics. His passion for genetics and, in particular, for inheritance problems, is what led him to think about calculation methods such as regression and correlation. Thus, the reflections that lead him to this field begin with a complicated (then) inheritance problem: the understanding of the force with which the characteristic of one generation of living beings manifests itself in the next. Initially, Galton approaches this problem by examining characteristics of pea seeds. Choose the pea because this species can self-fertilize: the daughter plants show genetic variations of the mother plants without the contribution of a second parent. In this way, Galton postpones the problem of statistically calculating genetic contributions from various sources. Galton's first idea about regression comes from a graph, a two-dimensional diagram, in which the sizes of the pea children were represented in front of those of the pea parents. Galton realized that the median diameter of the seed seeds for a particular diameter of the parent seed describes, approximately, a straight line with a positive slope and less than 1. This author uses the representation of his data to illustrate the basic foundations of the that statisticians continue to call regression. From here, with the errors of any incipient research process, Galton begins to build a whole theory that, mathematically, was later formalized by one of his disciples, Pearson.

So, the teaching objective of the experience developed has been a mixture of "problem-based learning" and "historical birth and development". Pragmatism and history. Mathematical modeling has been a posteriori. We invert the order, from practice to theory, from students to teacher: problems motivate, students think and propose solutions, and the teacher supervises and guides. With this we try to improve the understanding of the fundamentals and encourage the interest of the student by showing the various problems with which Galton, and other early researchers confronted and solved when they initiated the techniques that are so widely used today.

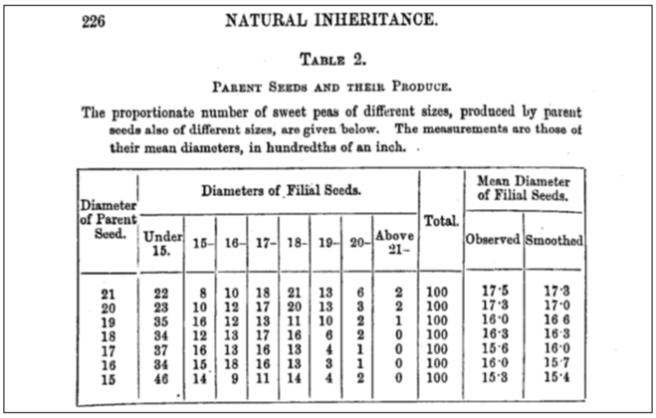
The experience has been developed in a group of about 80 students of the Degree in Business Administration and Management, in the subject Statistics I which constitutes an introductory course to statistics, and in which most of its content is related to techniques and Descriptive Statistics methods.

The path followed, then, has gone to the following address:

1. First, some biographical aspects of Sir Francis Galton are reviewed. Students are referred to the website recognized as official, about the life and work of this author: http://galton.org/. This circumstance allows a historical recount of the aspects of scientific and mathematical progress in the late nineteenth and early twentieth centuries. In this context, the scientific background with which Galton confronts the problem has been described and it has been explained how its mathematical deficiencies initially limited the analytical development of regression. Students also access the most pressing problems for science at the end of the 19th century, one of them being the genetic inheritance.

2. Employment in the classroom of some of the historical examples, with the same data as these pioneers. The first set of data that was offered to the stu-

dents is that which Galton worked on in his famous work Natural Inheritance (1894). In his four biographical volumes, Pearson describes the genesis of the discovery of the slope of regression (Pearson 1930). In 1875 Galton distributed packages of pea seeds among seven friends; each of them received seeds of uniform diameter (see also Galton 1894), but there were substantial differences between the different packages. Galton's friends collected the seeds of the new generation and returned it to him. The measurements of the diameters of this second generation crossed with those of the parents are collected by the author in the following table, this being the first one that is provided to the students:



226 NATURAL INHERITANCE.

TABLE 2.

PARENT SEEDS AND THEIR PRODUCE.

The proportionate number of sweet peas of different sizes, produced by parent seeds also of different sizes, are given below. The measurements are those of their mean diameters, in hundredths of an inch.

Diameter of Parent Seed.	Diameters of Filial Seeds.								Total.	Mean Diameter of Filial Seeds.	
	Under 15.	15–	16–	17–	18–	19–	20–	Above 21–		Observed	Smoothed
21	22	8	10	18	21	13	6	2	100	17·5	17·3
20	23	10	12	17	20	13	3	2	100	17·3	17·0
19	35	16	12	13	11	10	2	1	100	16·0	16 6
18	34	12	13	17	16	6	2	0	100	16·3	16·3
17	37	16	13	16	13	4	1	0	100	15·6	16·0
16	34	15	18	16	13	3	1	0	100	16·0	15·7
15	46	14	9	11	14	4	2	0	100	15·3	15·4

Figure 1: Image of the Table published by Galton on the diameter of parent pea seeds versus children pea seeds.

With the data from that table, we propose the construction of an Excel-type database, in which three columns are specified: the two variables to be related, diameter of the parent seed, those of the seed children, and as the third column the absolute frequency, that is, the number of times that a particular couple repeats.

Since the data corresponding to the child seeds are presented by Galton grouped in intervals, a "class mark" is selected for each one. An example of the construction of this database is the following table:

Table 1: Data in excel of the Table published by Galton on the diameter of seeds of pea parents versus seeds of pea children.

Seed diameter father	Seed diameter son	Frequency	Seed diameter father	Seed diameter son	Frequency
21	14,5	22	18	17,5	17
21	15,5	8	18	18,5	16
21	16,5	10	18	19,5	6
21	17,5	18	18	20,5	2
21	18,5	21	17	13,5	37
21	19,5	13	17	14,5	16
21	20,5	6	17	15,5	13
21	22,5	2	17	16,5	16
20	14,5	23	17	17,5	13
20	15,5	10	17	18,5	4
20	16,5	12	17	19,5	1
20	17,5	17	16	14,5	34
20	18,5	20	16	15,5	15
20	19,5	13	16	16,5	18
20	20,5	3	16	17,5	16
20	22,5	2	16	18,5	13
19	14,5	35	16	19,5	3
19	15,5	16	16	20,5	1
19	16,5	12	15	13,5	46
19	17,5	13	15	14,5	14
19	18,5	11	15	15,5	9
19	19,5	10	15	16,5	11
19	20,5	2	15	17,5	14
19	22,5	1	15	18,5	4
18	14,5	34	15	19,5	2
18	15,5	12			
18	16,5	13			

In a graph, Galton presented the diameters of the parental peas versus those of the children. As has been said, discover how the median diameters of the seed seeds for each specific diameter of the parent seed describes, approximately, a straight line with a positive slope and less than 1. Thus, naturally, it found a first regression line and also, a constant variability for all the series of a second character, for a given character of the first. Perhaps, the study of this simple special case was the best for the progress of correlational calculation, given the ease of understanding by a beginner. Therefore, following this process of conceptualization, the use of graphic methods is proposed first. We show the students the graphic representation of the dispersion diagram associated with this data. Excel allows this representation. In the classroom we have a computer for the teacher, with screen projection and, in turn, students come to class with a laptop mostly. Now, in order to facilitate the theoretical introduction of the concept, we recommend that you proceed to the standardization of both variables, leaving the two centered on the origin of coordinates (the two means are transformed into (0,0)) and both with equal variability (their standard deviations are made equal to 1), so that the differences in their magnitudes or scales do not interfere (masking or fattening) in the analysis of the possible relationship between both. Although we warn students that, in the case at hand, the two variables are very similar in terms of centrality and variability so, perhaps, their standardization would not have been necessary. With the typified variables we represent and, thus, visualize the relationship, inviting students to search for simple functions (straight lines) that are capable of reflecting as best as possible what the scatter diagram transmits. We accompany each point of the diagram of a numerical value that coincides with the corresponding absolute frequency. In some way, this value informs us of the weight that the associated point must have in the analysis of that relationship.

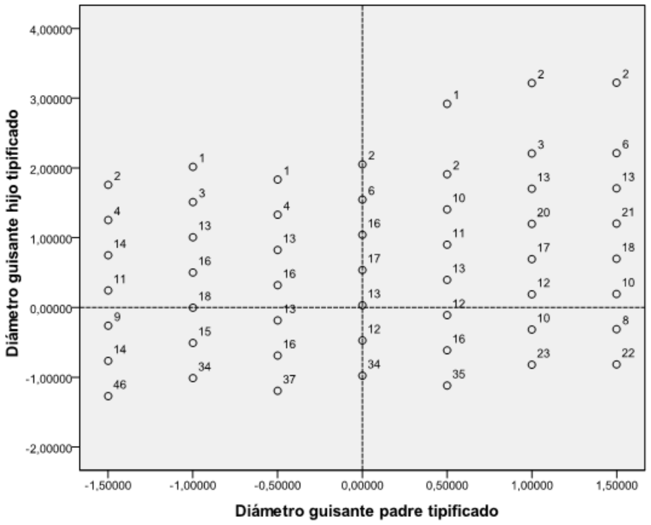


Figure 2: Dispersion diagram of the Galton data typified together with the absolute frequencies of the corresponding pairs.

The approximate line of a line interspersed between the points that try to approximate as much as possible to all of them, passing through the origin of coordinates since this is the "center" of the diagram and that takes into account the "weight" of each point in the graphic is what Galton tried as a first approximation and is what is presented before the vision of the data in the plane. Once each student has drawn his own line he is invited to calculate the slope of the same, easy calculation on the other hand as it is a quotient between opposite and contiguous distance. In this context, the student realizes two details: the straight line is increasing (the points located in the third quadrant have a lot of weight), but with a gentle slope (of course, much smaller than 1 that would correspond to the bisector of the 1st and 3rd quadrants). The first one informs us of a direct relation between both variables: in general, small diameters of the father correspond to small ones of the son and, on the other end, to large diameters of the parent, large diameters of the son seed. The second one makes us think that large variations in the parent diameters translate into smaller variations in the second generation, that is, values closer to the center in this second (what Galton called "regression to the mean", which gave rise to the term that nominates all this theory).

A second graphic proposal is proposed. The students already know the construction of the box diagrams. In this second graph, the median makes the paper that the media made in the previous one. We report that Galton's first attempts at constructing his regressions were using the median which, although he found it more intuitive, presents algebraic difficulties that impede the associated calculation developments.

In any case, this visualization corroborates the idea that was already forming. An interleaved line would pass through the origin of coordinates, it would have a positive slope, but it would be less than one:

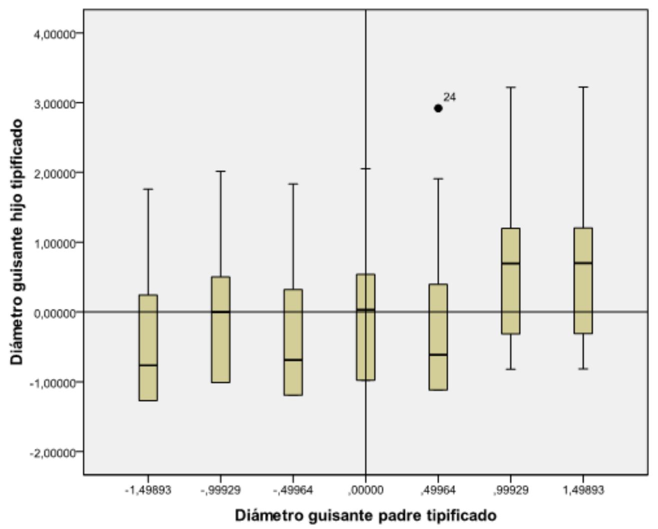


Figure 3: Box diagram of the Galton data typified.

3. Using a scientific calculator, first, and an Excel spreadsheet, afterwards, the students estimate the "slope" that, in some way, will define the possible relationship. The ease of the instruments allows to repeat different calculations and, even, to try different measurements, until finding the best approximations. So, we apply trial and error methods. In the approximations made by the students on this occasion we find slopes between 0.30 and 0.50. To that slope Galton called it r (regression).

4. It is the moment of formalization of the idea. Intuition makes us think that what is raised is an optimization problem: obtain the best line that represents the point cloud, that is, that minimizes the sum of the distances (squared) of the points of the scatter diagram to the line. The problem recalls one already solved when working with one-dimensional variables: minimization of the distances (squared) of the values of the distribution to a central value or, in other words, attempt to substitute the values of that distribution for a single value to represent them, a solution offered by the König Theorem (already known by the students at this stage of the development of the discipline) and which leads directly to the arithmetic mean as the representative and optimum value in the sense of distances. It only remains to make the jump to the two-dimensional case. Instead of one variable, we have two. Generalizing, the optimal solution is again an arithmetic mean, being in this case that of the resulting variable when constructing the cross products of the original variables. Therefore, the solution, the desired slope, is the average of the products crossed between both variables or, what Pearson called, "moment-product". We can write, then $r = \overline{XY}$, and the equation of the line that "adjusts" the cloud of points as best as possible is given by $y = rx$, where y represents the variable "diameter of typed children", that is, the variable effect, or variable to explain, or dependent, while x is that of the "diameters of the typed parents", variable cause, or explanatory, or independent. The line, as expected, passes through the origin of coordinates. For the standardized Galton data we obtain r = 0.346. Then, the adjusted line is represented between the point cloud:

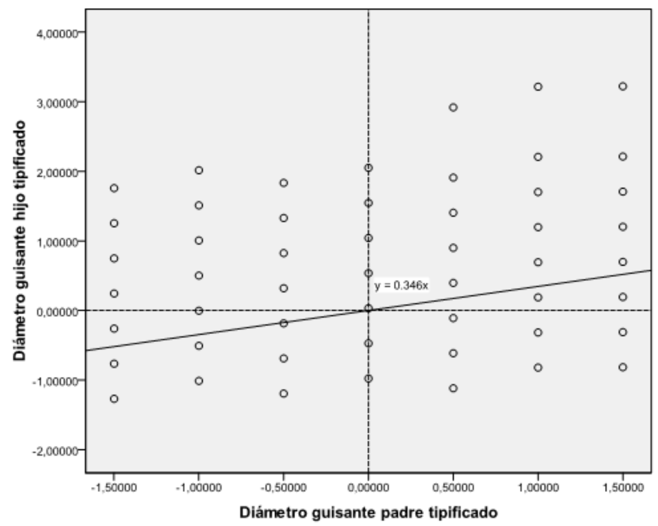


Figure 4: Dispersion diagram of the Galton data typified together with the straight line using the momentum-product as a slope.

With several simple examples, with few data, we show the effect of the different variability of x and y on the slope of the slope of the line. The objective is, from the slope initially constructed for the two variables with equal dispersion, to derive the resulting slopes in cases where the dispersion of x is greater than that of y, when both are equal (as in the case of previous data typified) and when the variability of x is less than that of y. The student captures how these variabilities, measured through their typical deviations, are making the line tilt (when $S_x > S_y$) or rise (when $S_x < S_y$). Then, with various calculation tests and relying on the previous optimization itself, we come to formulate the following equation $y = r(S_y / S_x)x$.

From this point, the student begins to distinguish the two fundamental parameters of the regression analysis: r, which we call correlation coefficient (degree of linear relationship between the two variables) and $b = r\frac{S_y}{S_x}$,

the slope of the line, where in the quotient the typical deviations of the two related variables. The three previous examples have been prepared so that they have the same value of r, so that the one that starts in this study captures that the change of slopes is due, in these cases, to the existing difference in the variability of both. We get the student to be aware of the difference between correlation and slope, and to accept the quotient $\frac{S_y}{S_x}$ as a corrector

or equalizer between both terms. By the way, the proportion is served: the slope is the standard deviation of y as the correlation coefficient is to that of x, or the ratio between slope and correlation coefficient is the same as between the standard deviation of y and that of x : $\frac{b}{r} = \frac{S_y}{S_x}$

The return to the original variables, undoing the typifications, brings out in a simple way the constant or independent term on the line.

5. Through Excel we calculate adjusted values and errors or residuals (differences between observed and adjusted). It is easy to check in the spreadsheet that the residuals add up to zero, that the compensatory effect between positive and negative errors is total with the adjusted line. But the errors exist and can be higher or lower depending on the dispersion level of the point cloud. The need for a measure of goodness of fit arises. It must be an aggregate, a summary measure of the errors. Since errors can be positive and negative, a quadratic measure avoids the possible compensatory effect of the sign. The sum of squares of errors or residuals is then proposed: SCR, which will be 0 when the adjustment is perfect and the greater the more dispersion there is in the diagram, the greater. One way to give security to work is to propose to the student the use of other values for the slope of the line, that is, other adjustments of the same, and the corresponding associated calculation of the SCR to show that, in any case, that aggregate of errors is always greater. The SCR is an absolute measure of the goodness of fit. The same, of course, depends on the scale used for the dependent variable. On the other hand, we can calculate the variability of said variable, through its variance, which we represent by VT (total variance), also the variability of the adjusted or explained values, VE (explained variance) and, finally, the variability of the errors or residuals, which is not more than the SCR converted into the average when divided by the sample size and that we represent by VR (residual variance). The calculations lead the student to check the following intuitive relationship: the fundamental basis of the calculation of the regression. It leads us to propose a second measure of goodness of fit, in this case relative and, therefore, useful to compare with other adjustments: "Proportion of variance explained". We call it the Determination Coefficient and its representation and calculation are defined in the following equality $R^2 = \frac{VE}{VT}$.

The student's surprise is capitalized when he verifies that this coefficient coincides with the square of r which, in turn, justifies the symbolism used to represent it.

6. The jump to multiple regression will be natural and, for the student, almost necessary when he is aware of the need to introduce more than one influencing factor in the objective variable. We illustrate how Galton realized, shortly after having collected and analyzed his data on peas, that the previous generation of immediate parents can also influence individual characteristics (Pearson, 1930). He points out that, even, certain characteristics skip one or more generations, occasionally; A man may be more like his grandfather than his father, in certain aspects. In an 1898 article in the journal Nature (cited in Pearson, 1930), Galton published an ingenious diagram that divided a square unit into successive smaller squares, each representing the diminishing influence of the previous generations of the ancestors. about the current individual. Galton came up with the germ of the idea of multiple regression. A characteristic or variable can be influenced not only by a single important cause, but by many causes of greater and lesser importance. Some of these causes may even overlap each other (that is, the explanatory variables are correlated with each other). In later publications Galton listed some mathematical formulas that picked up this same basic idea, but he was never able to develop a complete mathematical treatment of the subject:

"The somewhat complicated mathematics of multiple correlation, with its repeated appeals to the geometric notions of hyperspace, left him a closed room." (Pearson 1930, p.21)

However, Galton's conceptualization of the multiple influences of the ancestors on the characteristics of the individual of the present was completely parallel to the modern conception of multiple regression. As with the simple linear regression and the correlation coefficient, Galton put the imaginative preliminary work that Pearson later develops with a rigorous mathematical treatment. Pearson's subsequent work included the further development of multiple regression as well as innovative progress in other statistics. Then, using the formulation (although updated) and Pearson's examples, we introduce the student to multiple regression.

3. RESULTS AND DISCUSSIONS:

Once the experience is over, the professors who intervene in the same have three instruments to evaluate their result: the work developed by the students during the classes, the survey of the students and the qualification of the exam on this subject. In Spain, the qualification of an exam has a rank between 0 (minimum qualification) and 10 (maximum qualification). An exam is considered to have passed when its grade is greater than or equal to 5.

Other practical examples were proposed, all extracted from the history of this discipline. Some developed in class, with assistance and supervision of teachers. Students were asked, in addition to the appropriate calculations, the preparation of a conclusive report on what was extracted from these calculations. In this way, we take the opportunity to initiate them in the drafting of this type of report. The works, written in Word and sent by e-mail were rated on a scale of 0 to 10. The ratings of these works yielded the following results:

- The average score was 7.71 with a standard deviation of 2.03.
- The 50th percentile, or median, was 8.50, while the 75th percentile takes the value of 9.1, which gives us an idea of the majority of grades between outstanding and outstanding.
- A histogram of these ratings is shown in Figure 5.

The student survey was developed at the end of the experience. They were presented with a series of affirmations about which they manifested from their "total disagreement" to their "total agreement", on a Likert scale of five categories, which follow the indicated path. At the end of the survey, they were asked to make a global assessment of the experience through a rating, on a scale of 0 to 10.

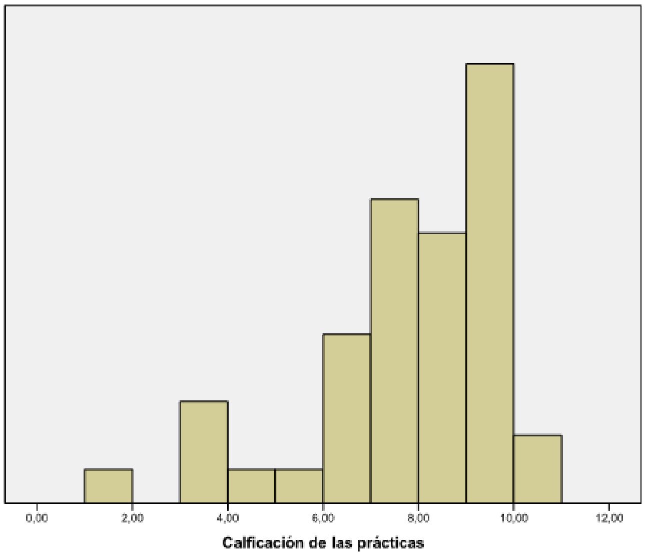


Figure 5: Histogram of the qualifications of the practices. Own elaboration based on student survey data

The following tables show percentage results of some of the statements made in the survey.

Affirmation: The historical context has helped me understand why the regression.

Table 2. Distribution of answers to question 1. Own elaboration based on student survey data.

Possible answers	Percentage
Neither agree nor disagree	2.9
In agreement	57.1
Totally agree	40.0

Affirmation: The experience developed has helped me to get loose in the handling of Excel.

Table 3. Distribution of answers to question 2. Own elaboration based on student survey data.

Possible answers	Percentage
In disagreement	5.7
Neither agree nor disagree	25.7
In agreement	57.1
Totally agree	11.4

Affirmation: The learning of Statistics is more motivated using historical contexts.

Table 4. Distribution of answers to question 3. Own elaboration based on student survey data.

Possible answers	Percentage
Neither agree nor disagree	11.4
In agreement	30.0
Totally agree	28.6

Regarding the qualification that the students give to this experience, we summarize it in the statistics that appear in the following table (qualification with rank between 0 and 10).

Table 5. Qualification of the experience by the students. Own elaboration based on student survey data.

Arithmetic average		7.7
Median		8.0
Mode		8.0
Standard deviation		0.7
Quartiles	25	7.0
	50	8.0
	75	8.0

Finally, we show what, perhaps, we are most interested in as teachers: knowing if the experience developed has positively contributed to the student's learning. The way in which we intend to evaluate is the written exam, similar in structure and contents to previous courses. The most important statistics related to the exam note corresponding to this subject are the following:

Table 6. Exam grades. Own elaboration based on student survey data.

Arithmetic average		6.04
Median		6.10
Mode		10.0
Standard deviation		2.39
Quantiles	25 33	3.90 5.00
	50 61	6.10 7.00
	75	7.70

Therefore, 67% of the students exceeded the subject, with almost 40% qualified with outstanding or outstanding. We have taken the grades corresponding to this same group, but from the previous year, and through a t test for independent samples, we compared the grades for those two consecutive years. We show results:

Table 7. Comparison of results of two consecutive years. Own elaboration based on student survey data.

Year	Arithmetic average	Standard deviation
2017	3,93	2,33
2018	6,04	2,39

Table 8. Test t for equality of means. Own elaboration based on student survey data.

Test of Levene for the equality of variances		Test T for equality of means				
F	p-value	t	gl	p-value (bilateral test)	Difference of means	Standard error of the difference
0.000	0.989	-4.131	84	0.000	-2.10	0.51

Assuming equal variability in the grades of one and another course (which confirms Levene's test), the difference between the average scores, of something more than 2 points, on a scale of 0 to 10, in favor of the grades of the last year (when the experience was carried out) is statistically significant ($p = 0.000$) according to this Student's t-test.

We are not so daring as to think that the difference in the average marks of both courses is due exclusively to the development or not of the commented experience. We are very aware, and our experience as teachers tells us so, that in each group and in each course many factors intervene, some known by the teacher and others not, that influence the exam grades. Therefore, it is difficult for us to assess, how the development of experience has weighed in that difference of notes. We think that the results presented in this section can give a vision, even if it is approximate, that could be valid to make a positive assessment of this proposal.

4. CONCLUSION:

The history of scientific progress is a good didactic argument. Confronting the student with the same problems with which the pioneering scientists laid the foundations for new theories is another way of motivating learning. If we combine a process of trial and error, intuitive, different approaches to the solution of the problem, given the possibilities that allows a spreadsheet, we can ensure the reinforcement and understanding of the subject taught. The theoretical formalism a posteriori. In short: intuition and calculation, observation and reflection. Route of a path that scientists of important stature followed at the time.

REFERENCES:

1. DUKE, J. D. (1978). Tables to Help Students Grasp Size Differences in Simple Correlations. Teaching of Psychology, 5, 219-221.
2. FITZPATRICK, P. J. (1960). Leading British Statisticians of the Nineteenth Century. Journal of the American Statistical Association, 55, 38-70.
3. GALTON, F. (1894). Natural Inheritance (5th ed.). New York, Macmillan and Company.
4. GOLDSTEIN, M. D., STRUBE, M. J. (1995). Understanding Correlations: Two Computer Exercises. Teaching of Psychology, 22, 205-206.
5. KARYLOWSKI, J. (1985). Regression Toward the Mean Effect: No Statistical Background Required. Teaching of Psychology, 12, 229-230.
6. PEARSON, E. S. (1938). Mathematical Statistics and Data Analysis (2nd ed.). Belmont, CA: Duxbury.
7. PEARSON, K. (1896). Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia. Philosophical Transactions of the Royal Society of London, 187, 253-318.
8. PEARSON, K. (1922). Francis Galton: A Centenary Appreciation. Cambridge University Press.
9. PEARSON, K. (1930). The Life, Letters and Labors of Francis Galton. Cambridge University Press.
10. STANTON, J. M. (2001). Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors. Journal of Statistics Education Vol 9, N. 3